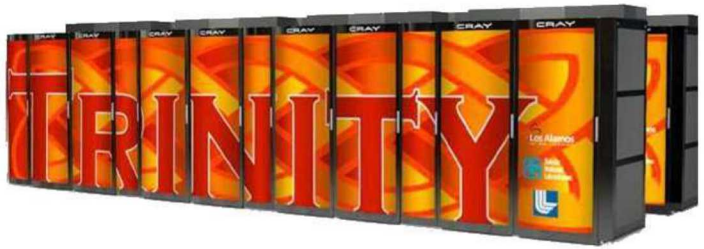
SAND2020-11026C

RaDD Runtimes: Radical and Different Distributed Runtimes with SmartNICs



PRESENTED BY

Ryan E. Grant, Whit Schonbein and Scott Levy

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

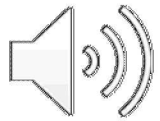
Images courtesy of: Sandia National Laboratories, Cray Inc.,

Why Now?

- Radically different systems software?

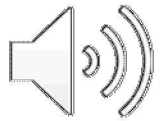
- New capable hardware types
  - Compute-capable SmartNICs
    - Bluefield, Stingray, Fungible DPUs

- New SmartNIC architectures
  - In Network Compute Assistance (INCA)
  - Allow deadline-free compute on NICs



What's different from the past?

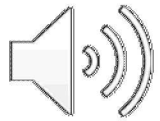
- Past attempts were slower than CPU cores
  - SmartNICs from early 2000s couldn't keep up with advances in CPU frequencies
- Past attempts had deadlines
  - Could only work on data for limited amount of time
  - Hard to program for; couldn't do enough to really make an impact



Is offloading radical?

- Just moving compute from CPU cores to network cores is simple
  - Reduces overhead and can increase performance

- Offloading is not enough for radical change
  - SmartNICs can enable entirely new applications and types of system software that can be independent

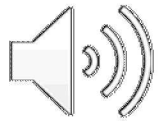


What is a SmartNIC?

- To understand the transformative nature of SmartNICs must define SmartNIC

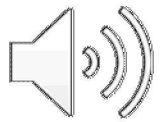
- Many definitions of SmartNICs are out there
  - Mostly marketing terms, not well defined

- Some companies use the term Data Processing Unit (DPU)
  - Really just a SmartNIC rebrand, may define a certain class of SmartNIC



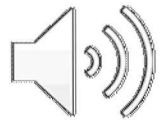
SmartNIC Classification

- Very simple classification (formal classification method coming in the future)
- Three basic classes of SmartNIC
- Class I: SmartNIC that just does basic network task offloads
  - Only does operations on network traffic, traditional offloads – not very smart



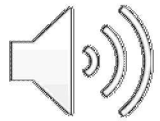
SmartNIC Classification

- Class II: SmartNIC that does general compute on data streams (sPIN)
  - Has to complete compute be deadline to free up resources for future incoming packets
  - Model: Unlimited incoming data, therefore limited compute
- Class III: SmartNIC that does general compute without deadline (INCA, bluefield)
  - Runs independent programs, not kernels on each packet
  - Model: Unlimited compute, only relies on some incoming data



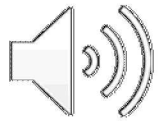
Radical Runtimes

- Only consider Class III SmartNICs for radical runtimes
  - Need unbounded compute times and complex applications on SmartNIC
- Significant compute capability
  - Similar to CPU cores, can even do some basic ML and much more complex compute with appropriate additional accelerators
- Assume we have “white” silicon equal to current NIC ASIC packages minus current NIC logic size
  - Package sizes driven by EE concerns, not logic needs
  - Spare space to build in compute “for free”



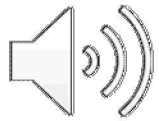
Radical Runtimes Overview

- Distributed Adaptive Resource Management
- System Tuning
  - Networks and node-level tuning
- Data Storage and Management
- Resilience
- Failure Prediction



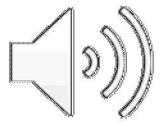
Distributed Adaptive Resource Management

- SmartNICs can enact adaptive resource management
- Control network resources amongst multiple actors
- Coordinate with other nodes on job-level resources management
- For elastic jobs, SmartNICs could even serve as job managers
  - E.g. SmartNICs running zookeeper in cloud



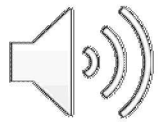
System Tuning

- Network tuning:
  - SmartNIC observes all traffic, so it can schedule and anticipate bandwidth needs and switch MTUs/small-large packet modes on the fly
  - Suggest times to send non-critical data when network is quiet
- System Tuning:
  - Let local tuning mechanism still operate
    - Close to performance counters on device
  - SmartNIC works as a coordination device between nodes
    - Coordinator for optimization frameworks like Intel's GEOPM



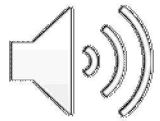
Data Storage and Management

- Expand Storage Hierarchy
  - SmartNICs have RAM/Storage
  - Extra level in hierarchy
  - Smart storage level with compute
- Could create Distributed Hash Tables (DHT)
  - Built in SmartNIC memory with SmartNIC management
  - DHT can facilitate data exchange between applications (e.g. simulation and visualization)
  - DHT persistence could be used with compute resources



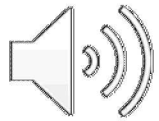
Data Storage and Management

- Coordinate I/O management
  - Schedule times to write data out to storage
  - Read in data speculatively for applications
  - Provide an additional level of storage in the hierarchy
  - Allow intelligent data movement to I/O to reduce application traffic interference
- Facilitate data movement
  - Simulation and viz/analysis
  - Move only the data that's needed
  - Pre-process data to determine what's needed where
  - Schedule data movement to meet deadlines with flexibility



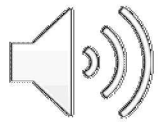
Resilience and Recovery

- SmartNICs can operate independently
  - Node crashes, SmartNIC stays up
  - For Kernel panic, many recovery options for data
- SmartNIC can manage communication state for recovery
  - Active notification of node failure – no more inferring nodes have crashed/gone away
  - Mitigates many zombie node situations
- Recover data from memory to backup nodes
- Without failures - manage checkpoints
  - Similar to I/O management



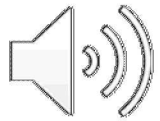
Failure Prediction

- SmartNICs can be used to perform more complicated failure detection schemes
  - No competition for compute resources with applications
- Easier to coordinate with whole system for distributed failure prediction
- Can use more information for failure prediction then just local conditions
  - Example: using surrounding node temperatures to determine future temperature spikes/drops that might increase chance of failure (heat cycling)



Takeaways

- Next Generation SmartNICs are creating exciting new possibilities
- Impact of non-deadline based models could be high
  - In pipeline like INCA or (mostly) out of pipeline like Bluefield
- Still much to learn about SmartNIC architecture design
- Many use cases still be to be thought of



Thank you

Questions?

Acknowledgments:

This work was funded through the Computational Systems and Software Environment sub-program of the Advanced Simulation and Computing Program funded by the National Nuclear Security Administration

References

- [1] Brian W Barrett, Ron Brightwell, Ryan Grant, Simon D Hammond and K Scott Hemmert. 2014. An evaluation of MPI message rate on hybrid-core processors. International Journal of High Performance Computing Applications 28, 4 (2014), 415–424.
- [3] Brian W. Barrett, Ron Brightwell, Ryan E. Grant, Scott Hemmert, Kevin Pedretti, Kyle Wheeler, Keith Underwood, Rolf Riesen, Torsten Hoefler, Arthur B. Maccabe, andTrammellHudson.2018. ThePortals4.2NetworkProgrammingInterface. Technical Report SAND2018-12790.
- [4] David E Bernholdt, Swen Boehm, George Bosilca, Manjunath Gorentla Venkata, Ryan E Grant, Thomas Naughton, Howard P Pritchard, Martin Schulz, and Geoffroy R Vallee. 2020. A survey of MPI usage in the US exascale computing project. Concurrency and Computation: Practice and Experience 32, 3 (2020), e4851.
- [5] Nanette J Boden, Danny Cohen, Robert E Felderman, Alan E. Kulawik, Charles L Seitz, Jakov N Seizovic, and Wen-King Su. 1995. Myrinet: A gigabit-per-second local area network. IEEE Micro 15, 1 (1995), 29–36.
- [6] Pat Bosshart, Dan Daly, Glen Gibb, Martin Izzard, Nick McKeown, Jennifer Rexford, Cole Schlesinger, Dan Talayco, Amin Vahdat, George Varghese, and David Walker. 2014. P4: Programming Protocol-independent Packet Processors. SIGCOMM Comput. Commun. Rev. 44, 3 (July 2014), 87–95. https://doi.org/10.1145/2656877.2656890
- [7] Ron Brightwell, Kevin T Pedretti, Keith D Underwood, and Trammell Hudson. 2006. SeaStar interconnect: Balanced bandwidth for scalable performance. IEEE Micro 26, 3 (2006), 41–57.
- [8] Broadcom. 2019. Stingray SmartNIC. Retrieved 2019-10-01 from https://www.broadcom.com/products/ethernet-connectivity/smartnic/ps225
- [9] Darius Buntinas. 2012. Scalable distributed consensus to support MPI fault tolerance. In 2012 IEEE 26th International Parallel and Distributed Processing Symposium. IEEE, 1240–1249.
- [10] Darius Buntinas, Dhabaleswar K. Panda, and Ponnuswamy Sadayappan. 2001. Fast NIC-based barrier over Myrinet/GM. In Proceedings 15th Inter- national Parallel and Distributed Processing Symposium. IPDPS 2001. 52–59. https://doi.org/10.1109/IPDPS.2001.924993
- [11] Barbara Chapman, Tony Curtis, Swaroop Pophale, Stephen Poole, Jeff Kuehn, Chuck Koelbel, and Lauren Smith. 2010. Introducing OpenSHMEM: SHMEM for the PGAS community. In Proceedings of the Fourth Conference on Partitioned Global Address Space Programming Model. 1–3.
- [12] Christopher L Chappell and James Mitchell.2012.Packet processing in switched fabric networks. PatentNo.8285907,FiledDecember10th.,2004,IssuedOctober 9th., 2012.
- [13] Dennis Dalessandro, Ananth Devulapalli, and Pete Wyckoff. 2005. Design and implementation of the iWARP protocol in software. In Proceedings of the 17th IASTED International Conference on Parallel and Distributed Computing and Systems. Phoenix, Arizona, 471–476.
- [14] DennisDalessandro,PeteWyckoff,andGaryMontry.2006.Initialperformance evaluation of the neteffect 10 gigabit iwarp adapter. In 2006 IEEE International Conference on Cluster Computing. IEEE, 1–7.
- [15] Huynh Tu Dang, Daniele Sciascia, Marco Canini, Fernando Pedone ,and Robert Soulé. 2015. NetPaxos: Consensus at Network Speed. In Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research (SOSR '15). ACM, New York, NY, USA, 5:1–5:7. https://doi.org/10.1145/2774993.2774999 event-place: Santa Clara, California.
- [16] S. Derradji, T. Palfer-Sollier, J. P. Panziera, A. Poudes, and F. W. Atos. 2015. The BXI Interconnect Architecture. In 2015 IEEE 23rd Annual Symposium on High-PerformanceInterconnects.18–25. https://doi.org/10.1109/HOTI.2015.15
- [17] Ciprian Docan, Manish Parashar, and Scott Klasky. 2012. Dataspaces: an interaction and coordination framework for coupled simulation workflows. Cluster Computing 15, 2 (2012), 163–181.
- [18] MDosanjh,WSchonbein,REGrant,PGBridges,SMGazimirsaeed,andAAfsahi. 2019. Fuzzy Matching: Hardware Accelerated MPI Communication Middleware. In 2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). 210–220.
- [19] Jonathan Eastep,Steve Sylvester,Christopher Cantalupo,Brad Geltz,Federico Ardanaz, Asma Al-Rawi, Kelly Livingston, Fuat Keceli, Matthias Maiterth, and Siddhartha Jana. 2017. Global extensible open power manager: A vehicle for hpc community collaboration on co-designed energy management solutions. In International Supercomputing Conference. Springer, 394–412.
- [20] Christian Engelmann, Geoffroy R Vallee, Thomas Naughton, and Stephen L Scott. 2009. Proactive fault tolerance using preemptive migration. In 2009 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing. IEEE, 252–257.
- [21] Daniel Firestone, Andrew Putnam, Sambhrama Mundkur, Derek Chiou, Alireza Dabagh, Mike Andrewartha, Hari Angepat, Vivek Bhanu, Adrian Caulfield, Eric Chung, et al. 2018. Azure accelerated networking: SmartNICs in the public cloud. In 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18). 51–66.
- [22] Mario Flajslik, James Dinan, and Keith D Underwood. 2016. Mitigating MPI message matching misery. In International conference on high performance computing. Springer, 281–299.
- [23] Ana Gainaru,Franck Cappello, Marc Snir, and William Kramer.2013. Failure prediction for HPC systems and applications: Current situation and open issues. The International journal of high performance computing applications 27, 3 (2013), 273–282.
- [24] Balazs Gerofi, Masamichi Takagi, Atsushi Hori, Gou Nakamura, Tomoki Shirasawa, and Yutaka Ishikawa. 2016. On the scalability, performance isolation and device driver transparency of the IHK/McKernel hybrid lightweight kernel. In 2016 IEEE Inter-national Parallel and Distributed Processing Symposium (IPDPS). IEEE, 1041–1050.

References

- [25] René Glebke, Johannes Krude, Ike Kunze, Jan Rüth, Felix Senger, and Klaus Wehrle. 2019. Towards Executing Computer Vision Functionality on Programmable Network Devices. In Proceedings of the 1st ACM CoNEXT Workshop on Emerging In-Network Computing Paradigms (ENCP '19). Association for Computing Machinery,NewYork,NY,USA,15–20. https://doi.org/10.1145/3359993.3366646 event-place: Orlando, FL, USA.
- [26] Richard LG raham,Devendar Bureddy, Pak Lui,Hal Rosenstock,Gilad Shainer,Gil Bloch, Dror Goldenerg, Mike Dubman, Sasha Kotchubievsky, Vladimir Koushnir, et al. 2016. Scalable hierarchical aggregation protocol (SHArP): a hardware architecture for efficient data reduction. In Proceedings of the First Workshop on Optimization of Communication in HPC. IEEE Press, 1–10.
- [27] Richard L Graham, Steve Poole, Pavel Shamis, Gil Bloch, Noam Bloch, Hillel Chapman, Michael Kagan, Ariel Shahar, Ishai Rabinovitz, and Gilad Shainer. 2010. Overlapping computation and communication: Barrier algorithms and ConnectX- 2 CORE-Direct capabilities. In Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on. IEEE, 1–8.
- [28] Ryan E Grant, Mohammad J Rashti, Ahmad Afsahi,and Pavan Balaji.2011.RDMA capable iWARP over datagrams. In Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International. IEEE, 628–639.
- [29] Taylor Liles Groves, Ryan E Grant, Aaron Gonzales, and Dorian Arnold. 2017. Unraveling Network-Induced Memory Contention: Deeper Insights with Machine Learning. IEEE Transactions on Parallel and Distributed Systems 29, 8 (2017), 1907–1922.
- [30] K. Scott Hemmert,Brian Barrett,and Keith D. Underwood.2010.UsingTriggered Operations to Offload Collective Communication Operations. In Recent Advances in the Message Passing Interface (Lecture Notes in Computer Science). Springer, Berlin, Heidelberg, 249–256. https://doi.org/10.1007/978-3-642-15646-5_26
- [31] Nathan Hjelm,Matthew GF Dosanjh,Ryan E Grant,Taylor Groves,Patrick Bridges, and Dorian Arnold. 2018. Improving MPI multi-threaded RMA communication performance. In Proceedings of the 47th International Conference on Parallel Processing. 1–11.
- [32] Torsten Hoefler, Salvatore Di Girolamo, Konstantin Taranov, Ryan E. Grant, andRonBrightwell.2017. sPIN:High-performanceStreamingProcessingIn the Network. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '17). ACM, New York, NY, USA, 59:1–59:16. https://doi.org/10.1145/3126908.3126970
- [33] Dewan Ibtesham, Dorian Arnold,Kurt B Ferreira, and Patrick G Bridges.2011. On the viability of checkpoint compression for extreme scale fault tolerance. In European Conference on Parallel Processing. Springer, 302–311.
- [34] Humaira Kamal and Alan Wagner. 2010. FG-MPI: Fine-grain MPI for multicore and clusters. In 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW). IEEE, 1–8.
- [35] Antoine Kaufmann, SImon Peter, Naveen Kr. Sharma, Thomas Anderson, and Arvind Krishnamurthy. 2016. High Performance Packet Processing with FlexNIC. In Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '16). ACM, New York, NY,USA,67–81. https://doi.org/10.1145/2872362.2872367
- [36] Gokcen Kestor, Roberto Gioiosa, Darren J Kerbyson, and Adolfy Hoisie. 2013. Quantifying the energy cost of data movement in scientific applications. In 2013 IEEE international symposium on workload characterization (IISWC). IEEE, 56–65.
- [37] D. Brian Larkins,John Snyder,and James Dinan.2018.EfficientRuntimeSupport for a Partitioned Global Logical Address Space. In ICPP 2018: 47th International Conference on Parallel Processing. ACM, Eugune, Oregon.
- [38] Jay Lofstead,John Mitchell,and Enze Chen.2020.StitchItUp:UsingProgressive Data Storage to Scale Science. In 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE, 52–61.
- [39] Mellanox. 2018. Mellanox BlueField SmartNIC. Retrieved 2019-10-01 from https://www.mellanox.com/products/bluefield-overview
- [40] Mellanox. 2018. Understanding MPI Tag Matching and Rendezvous Offloads (ConnectX-5). https://community.mellanox.com/s/article/ understanding- mpi- tag- matching- and- rendezvous- offloads- - connectx- 5- x
- [41] Netronome. [n. d.]. Agilio CX SmartNICs. https://www.netronome.com/ products/agilio- cx/
- [42] Fabrizio Petrini, Wu-chun Feng, Adolfy Hoisie, Salvador Coll, and Eitan Frachtenberg. 2002. The Quadrics network: High-performance clustering technology. IEEE Micro 22, 1 (2002), 46–57.
- [43] James S Plank,Jian Xu,and Robert HB Netzer.1995.Compresseddifferences:An algorithm for fast incremental checkpointing. Technical Report. Citeseer.

References

- [44] Mohammad J Rashti, Ryan E Grant, Ahmad Afsahi, and Pavan Balaji.2010.iWARP redefined: Scalable connectionless communication over high-speed Ethernet. In High Performance Computing (HiPC), 2010 International Conference on. IEEE, 1–10.
- [45] Timo Schneider, Torsten Hoefler, Ryan E Grant, Brian W Barrett, and Ron Brightwell. 2013. Protocols for fully offloaded collective operations on accelerated network adapters. In Parallel Processing (ICPP), 2013 42nd International Conference on. IEEE, 593–602.
- [46] Whit Schonbein, Matthew GF Dosanjh, Ryan E Grant, and Patrick G Bridges.2018. Measuring multithreaded message matching misery. In European Conference on Parallel Processing. Springer, 480–491.
- [47] Whit Schonbein, Ryan E Grant, Matthew GF Dosanjh, and Dorian Arnold.2019. INCA: in-network compute assistance. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 54.
- [48] Min Si, Antonio J Peña, Pavan Balaji, Masamichi Takagi,and Yutaka Ishikawa. 2014. MT-MPI: multithreaded MPI for many-core environments. In Proceedings of the 28th ACM international conference on Supercomputing. 125–134.
- [49] Min Si, Antonio J Peña, Jeff Hammond, Pavan Balaji, Masamichi Takagi, and Yutaka Ishikawa. 2015. Casper: An asynchronous progress model for MPI RMA on many-core architectures. In 2015 IEEE International Parallel and Distributed Processing Symposium. IEEE, 665–676.
- [50] Pradeep Sindhu, Jean-Marc Frailong, Wael Noureddine, Felix A Marti, Deepak Goel, Rajan Goyal, and Bertrand Serlet. 2019. Data processing unit for stream processing. USPatentApp.16/031,945.
- [51] Giuseppe Siracusano and Roberto Bifulco.2017.Is it a SmartNIC or a Key-Value Store? Both! In Proceedings of the SIGCOMM Posters and Demos. 138–140.
- [52] Craig Ulmer, Shyamali Mukherjee, Gary Templet, Scott Levy, Jay Lofstead, Patrick Widener, Todd Kordenbrock, and Margaret Lawson. 2018. Faodel: Data management for next-generation application workflows. In Proceedings of the 9th Workshop on Scientific Cloud Computing. 1–6.
- [53] K.D.Underwood,J.Coffman,R.Larsen,K.S.Hemmert,B.W.Barrett,R.Brightwell, and M. Levenhagen. 2011. Enabling Flexible Collective Communication Offload with Triggered Operations. In 2011 IEEE 19th Annual Symposium on High PerformanceInterconnects.35–42. https://doi.org/10.1109/HOTI.2011.15
- [54] S Van Doren. 2019. Compute Express Link, Stephen Van Doren (Intel). In 2019 IEEE Symposium on High-Performance Interconnects (HOTI). IEEE, 18–18.
- [55] A. Wagner, Hyun-Wook Jin, D. K. Panda, and R. Riesen. 2004. NIC-based offload of dynamic user-defined modules for Myrinet clusters. In 2004 IEEE International Conference on Cluster Computing (IEEE Cat. No.04EX935). 205–214. https://doi.org/10.1109/CLUSTR.2004.1392618
- [56] Felix Zahn and HolgerFröning.2020.OnNetworkLocalityinMPI-BasedHPC Applications. In 49th International Conference on Parallel Processing-ICPP. 1–10.
- [57] Judicael A Zounmevo,Xin Zhao,Pavan Balaji,William Gropp,and Ahmad Afsahi. 2014. Nonblocking epochs in MPI one-sided communication. In SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 475–486.